Multiclass Classification of Breast Cancer in Whole-Slide Images

Scotty Kwok(✉)

Seek AI Limited, Hong Kong, China
scottykwok@gmail.com

Abstract. Breast cancer is one of the leading cause of cancer-related death worldwide. During the diagnosis of breast cancer, the histopathological assessment of Haemotoxylin and Eosin(H&E) stained slides provides important clinical values. By applying computer-aid diagnosis on whole-slide image(WSI), the efficiency and consistency of such assessment could be improved. In this paper, we propose a deep learning-based framework that classifies H&E stained WSIs into regions of normal tissue, benign lesion, in-situ carcinoma and invasive carcinoma. The framework utilizes both microscopy images and WSIs to train a patch classifier in two stages. The underlying classifier is based on Inception-Resnet-v2. This framework won both parts of the *ICIAR2018 Grand Challenge on Breast Cancer Histology Images* [4] competition, achieved a part A multiclass accuracy of 87% and part B score of 0.6929.

Keywords: Breast cancer · Deep learning · Whole-Slide Images
Multiclass classification

1 Introduction

Breast cancer is one of the leading cause of cancer-related death worldwide. According to the estimation of American Cancer Society, among US women in 2017, there will be an estimated 252,710 new cases of invasive breast cancer, 63,410 new cases of breast carcinoma in situ, and 40,610 breast cancer deaths [3]. The diagnosis of breast cancer involves the histopathological assessment of Haemotoxylin and Eosin (H&E) stained sections under microscope. The assessment result provides the basis for clinical treatment and management decisions, which significantly impact the mortality and quality of life of patients. Nevertheless, this manual assessment task is challenging due to the following reasons: (1) this task required experienced pathologists, (2) this task is tedious and time consuming, and (3) the result is subjected to variability in inter-rater and/or intra-rater concordance [6,8,11]. Computer-aided diagnosis (CAD) is therefore an appealing option for tackling these problems.

A. Campilho et al. (Eds.): ICIAR 2018, LNCS 10882, pp. 931–940, 2018.
https://doi.org/10.1007/978-3-319-93000-8_106

2 Related Work

Among the many CAD techniques, studies have shown that CNN-based analysis outperformed other methods in various pathological classification tasks. Specifically for studies related to breast cancer histopathology, researchers have published related works based on: BreakHis [13], Camelyon16 [1], Camelyon17 [2] and the enriched Bioimaging 2015 dataset [5]. For BreakHis, a recent study by Habibzadeh et al. [7] reported a binary classification accuracy of 98.7% using ResNet-152. For Camelyon16, the winning team, Wang et al. [16], achieved an area under the receiver operating curve (AUC) of 0.925 by using GoogleLeNet to detect metastases in lymph nodes. Later in Camelyon17, the winning team, Zhong et al. [16], achieved a Kappa score of 0.8958 in classifying the pN-stage of patients by using Resnet-101 and spatial pyramid pooling.

Despite most of studies focused on binary classification (normal vs tumor), multiclass classification actually offer more clinical values for an informed decision. The study of Araújo et al. [5] addressed this issue by enriching the Bioimaging 2015 dataset and using a CNN-based approach to classify histology images into four classes: normal, benign, in-situ carcinoma and invasive carcinoma. The authors achieved a state-of-the-art multiclass accuracy of 77.8% and binary accuracy of 83.3%.

Built on top of the enriched Bioimaging 2015 dataset, the ICIAR 2018 Grand Challenge on Breast Cancer Histology Images (BACH2018) dataset further enriched the data with more microscopy images and added WSI into the collection. In this paper, we will present the framework that we used to participate in BACH2018, which has achieved promising results in the challenge. The schematic overview of the framework is in Fig. 1.

3 Materials

3.1 Part a - Microscopy Images

The train set consists of 400 microscopy images of size 2048×1536 pixels, with pixel scale $0.42\,\mu m$. The images were evenly sampled from each classes: Normal (100), Benign (100), In-situ carcinoma (100) and Invasive carcinoma (100). This dataset is an extension of the one used by Araújo et al. [5], the data collection methodology was explained in details in their article. The patient-wise origin of each microscopy image is only partially available due to the anonymization process. The test set consists of another 100 microscopy images of the same scale. The ground truth is hidden.

3.2 Part B - Whole-Slide Images

The train set consists of 10 WSIs in various sizes (e.g. 42113×62625 pixels), with pixel scale $0.467\,\mu m$. These whole-slide images are high resolution images containing the entire sampled tissue. The annotation was prepared by two medical

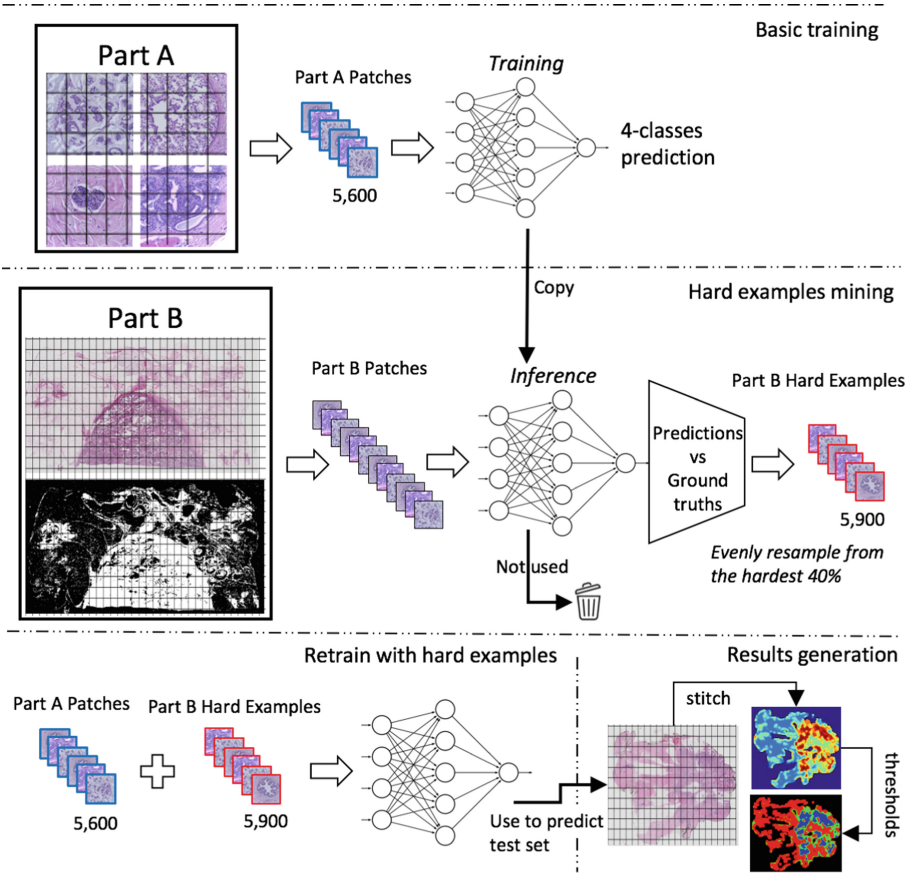
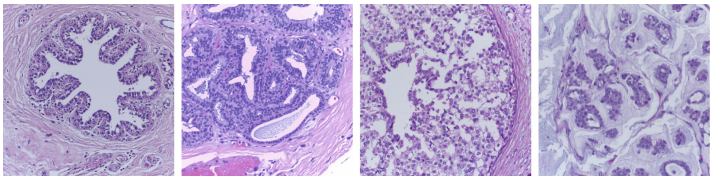


Fig. 1. Overview of the framework



(a) Normal (b) Benign (c) In-situ (d) Invasive

Fig. 2. Sample Part A Patches

experts and images where there was disagreement were discarded. The ground truth annotations are multiple regions that were labelled as: Benign, In-situ carcinoma or Invasive carcinoma. The patient-wise origin of the whole-slide images are fully available. (Note the train set also contains another 20 whole-slide images without ground truth, these 20 slides were not used in our approach). The test set consists another 10 WSIs of the same scale. The ground truth is hidden.

4 Methods

4.1 Patch Extraction from Microscopy Images and Augmentations

Patches were cropped from each of the images in Part A, using a patch size of 1495×1495 pixels and stride of 99 pixels. The 400 microscopy images were cropped into 5,600 patches (examples in Fig. 2). These patches were then resized to 299×299 pixels. To utilize the rotational symmetry, random vertical/horizontal flipping and rotation of 90, 180, 270° were applied. To combat the color variation of H&E stain, random HSV color space augmentations were applied.

4.2 Choice of CNN

Given the limited data size and high model capacity of CNN, we postulated that those existing CNN architectures are sufficient to handle this task. Four existing CNN architectures (VGG19 [12], Inception-v3 [15], Inception-v4 and Inception-Resnet-v2 [14]) were selected and tested empirically. The test involved splitting the Part A microscopy images into train set(75%) and held-out set(25%). Patches were extracted from each sets. The CNNs were trained and tuned to optimal accuracy, and patches from held-out set were used to evaluate the true predictive power of the CNNs on unseen data. Based on the test results in Table 1, Inception-Resnet-v2 outperformed others and was therefore chosen.

Table 1. Accuracy of different models

Model	Four-classes accuracy	Binary accuracy
VGG19	0.70	0.81
Inception-v3	0.74	0.85
Inception-v4	0.71	0.82
Inception-Resnet-v2	0.79	0.91

4.3 Basic Training

In technical details: the top layers of Inception-Resnet-v2 were replaced by a fully connected layer with 2048 units, followed by a dropout layer with 50% dropout rate, and a fully connected output layer with Sigmoid as activation function. The output probabilities need to be normalized to sum to unity due to the use of Sigmoid. The network were initialized using ImageNet pre-trained weights. The back propagation was performed by Stochastic Gradient Descent with a constant learning rate(0.001), Nesterov momentum (0.9), batch size(64) and categorical cross-entropy as the objective function. Note only Part A patches were used in this training. Using a machine with two GPUs (GeForce GTX 1080 Ti), the model converged to its optimal accuracy within 25 epochs, in $<$35 mins.

4.4 Patch Extraction from WSIs

The Part B WSIs need to be converted to patches before they can be used. The conversion began with our customized foreground extraction. Unlike many prior works, where Otsu thresholding [10] or gray value thresholding [9] were used, our extraction method made use of the color characteristics of H&E stain to threshold tissue regions. In our method, WSI was down-sampled and converted from RGB to CIE L*a*b* color space. The mean intensity of the a* channel were then computed. And by applying a binary threshold on the a* channel, all the pixels that were above mean by 10% became the foreground. A sample result was shown in Fig. 3. The rationale of this method is based on the fact that H&E stained tissues are predominantly red/magenta in color, whereas the a* channel is a good approximation of how red/magenta a pixel is.

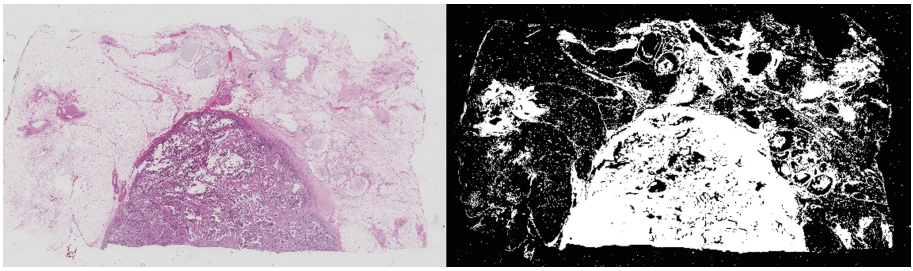


Fig. 3. The original WSI and the computed foreground mask (Color figure online)

Next, the WSIs were then cropped into patches of the same scale as that was done in Part A. Patches with less than 5% foreground pixels were considered as empty and discarded. The coordinates of the patches were stored in file for the later use during heatmap stitching.

Lastly, the WSIs ground truth annotations were converted into patch-wise class labels by the following method: each class were mapped to a pixel value based on its invasiveness, that is Normal = 0, Benign = 1, In-Situ Carcinoma = 2 and Invasive Carcinoma = 3. The patch-wise invasiveness was then be computed by taking the mean overall all the pixels in the patch. The mean invasiveness was then rounded to the nearest class and became the patch-wise class label.

4.5 Hard Examples Mining

The patch classifier (we trained earlier using Part A patches) was then used to predict Part B patches. By comparing the prediction verse the ground truth, the difficulty of each patch can be quantified. The patch difficulty was computed by: the absolute class distance between the ground truth class and the predicted class, multiplied by the predicted probability. For example: if given an invasive carcinoma patch, the classifier predicted 90% chance benign, then the difficulty

of this patch is: abs(3 − 1) * 0.9 = 1.8. This value enabled us to sort the patches according to difficulties.

To further narrow down the selection, patches with less than 70% foreground pixels were excluded. Finally, patches were sorted by difficulty and the top 40 percentile were sampled as candidates. The resultant number of patches were imbalanced: Normal (28,000), Benign (4,500), In-Situ Carcinoma (1,500) and Invasive Carcinoma (19,700). In order to re-balance the classes, a total of 5,900 patches were evenly sampled from each classes to become our final hard examples collection.

4.6 Retrain with Hard Examples

The patch classifier was retrained by combing Part A patches (5,600) and Part B hard examples (5,900), using the same CNN architecture and hyper-parameters. The model converged to its optimal accuracy within 15 epochs, in <40 min.

4.7 Results Generation

The test set was converted to patches in the same way and fed through the patch classifier to obtain predictions, except that image augmentation was disabled during inference.

Part A results are class labels. They were generated by averaging the patch-wise predictions in each images.

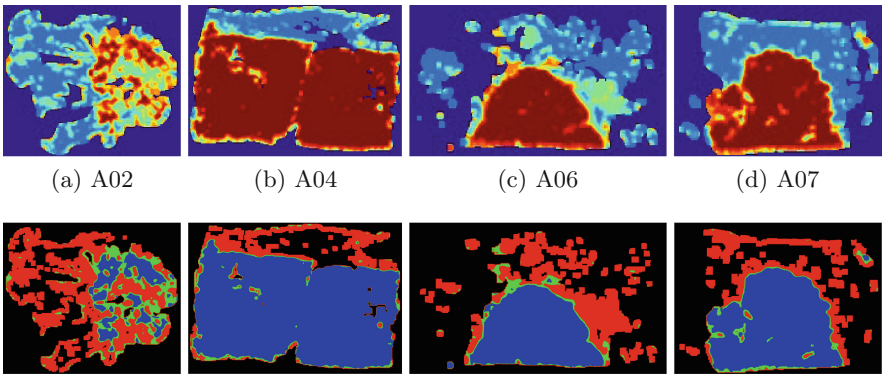
Part B results are color-coded class maps. They were generated by two steps: (1) First a heatmap was computed by stitching all the patch-wise predictions into one single image buffer, the pixel intensity were then normalized to a value between 0 and 1. The higher the value, the more likely that the pixel was invasive carcinoma. (2) Then in order to quantize the pixel intensity into four classes, the optimal thresholds were probed empirically. In our implementation, the thresholds for Normal, Benign, In-situ Carcinoma and Invasive Carcinoma were: 0, 0.35, 0.7, 0.75 respectively. Note when compared to the default thresholds (0, 0.25, 0.5, 0.75), our thresholds were biased towards predicting more Normal/Benign and less In-situ Carcinoma (Fig. 4).

5 Results and Analysis

5.1 Part a Results

The primary evaluation metric for Part A is multiclass accuracy. Our approach achieved a multi-classes accuracy of **87%** when presented with the test set. This is a **9%** improvement comparing to the previous best result reported by Araújo et al. [5], where their CNN+SVM approach achieved a multiclass accuracy of 77.8%.

Together with another team (Chennamsetty et. al.) that also achieved 87% accuracy in the challenge, we won the first place in Part A of the competition.



(a) A02 (b) A04 (c) A06 (d) A07

(e) A02 (f) A04 (g) A06 (h) A07

Fig. 4. (a)−(d) are the heatmaps generated by stitching patch-wise predictions into single image. (e)−(h) are the color-coded class map generated by applying thresholds to the heatmaps

The other evaluation metrics were not available because the test set ground truth were not disclosed at the time of this writing. The binary accuracy and the root-mean-square error should provide further insight to differentiate the model performances.

5.2 Part B Results

The primary evaluation metric for Part B is based on the following score, s:

$$s = 1 - \frac{\sum_{i=1}^{N} |\text{pred}_i - \text{gt}_i|}{\sum N_{i=1} max(|\text{gt}_i - 0|, |\text{gt}_i - 3|) \times [1 - (1 - \text{pred}_{i,bin})(1 - \text{gt}_{i,bin})]}$$

where "pred" is the predicted class (0, 1, 2 or 3), and "gt" is the ground truth class, i is the linear index of a pixel in the image, N is the total number of pixels in the image and bin is the binary value.

Our approach achieved a score of **0.6929** when presented with the test set. Again, it won the first place in Part B of the competition. To our knowledge, there were no prior studies for comparison, but our score outperformed the second (**0.5527**) and the third (**0.5230**) team by a large margin.

5.3 Part B Analysis

The evaluation methods and results of Part B deserved a more detailed analysis. Unfortunately, the test set ground truth were not disclosed at the time of this writing, so here we present the analysis using the train set WSIs (A01 to A10).

Firstly, inspection of the ground truth revealed that some annotated regions enclosed considerable amount of empty spaces (Fig. 5). This was inevitable

because (1) human annotations were high level, and (2) some of the histological structure enclosed empty spaces by itself. These empty spaces were not a concern for human interpretation, but for the evaluation of model performance, empty spaces could distort the true score. A simple and effective remedy was to exclude all the empty spaces in both the ground truth and the predictions. This can be done by making use of the foreground masks we obtained earlier, to mask out all the empty pixels during evaluation. Hence in our analysis below, we present two sets of scores: the basic scores and the scores that excluded empty spaces.

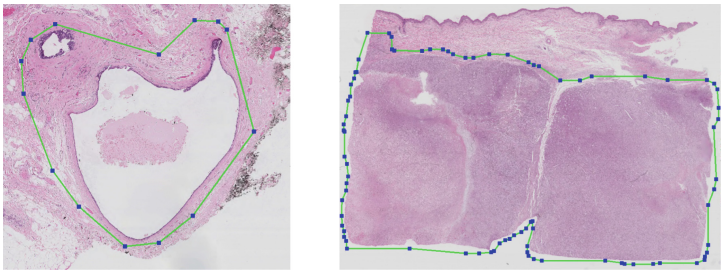


Fig. 5. Some annotated region enclosed considerable amount of empty pixels

Secondly, we established a baseline prediction by blindly predicting all pixels as Benign. This baseline strategy was based on the fact that the evaluation score does not weight the four classes equally. According to the organizers, the evaluation score was designed to penalize predictions that are farther from the ground truth value. And in the denominator, cases in which the prediction and ground truth are both 0 (normal class) are not counted, this was to avoid over-evaluating the correct predictions of normal cases. We tested this metric statistically and found that if the four classes are equally likely, then the expected score of predicting Normal, Benign, In-situ Carcinoma and Invasive Carcinoma are 0.14, 0.60, 0.60 and 0.40 respectively. In other words, the dominant strategy was to at least predict a pixel as Benign. Hence in our analysis below, we enriched the analysis by comparing our scores with this baseline.

Finally, listed in Table 2a are the scores of individual WSI predicted using the baseline verse our framework. Listed in Table 2b are the same set of scores but with empty spaces excluded in the evaluation. The bottom rows are the average scores.

In Table 2a, the average score of the baseline is 0.60. This suggests that any model that scores below 0.60 is no better than a blind prediction. Our approach achieved an average score of 0.75.

The individual scores of each WSIs in Table 2a reveal that our approach underperformed the baseline in some conditions, such as in slide A03 ($0.66 > 0.63$) and A08 ($0.66 > 0.57$). By inspection, we found that our classifier was weak at detecting the in-situ carcinoma regions in A03, and the small invasive carcinoma regions of A08.

Table 2. Part B analysis

WSI	Baseline	Our framework
A01	0.65	0.68
A02	0.60	0.71
A03	0.66	0.63
A04	0.44	0.92
A05	0.61	0.82
A06	0.61	0.85
A07	0.54	0.87
A08	0.66	0.57
A09	0.62	0.73
A10	0.62	0.74
Avg. Score	0.60	0.75

(a) Scores

WSI	Baseline	Our framework
A01	0.61	0.70
A02	0.52	0.75
A03	0.64	0.65
A04	0.37	0.67
A05	0.53	0.86
A06	0.51	0.90
A07	0.45	0.91
A08	0.65	0.56
A09	0.52	0.78
A10	0.53	0.79
Avg. Score	0.53	0.79

(b) Scores excl. empty spaces

The differences in average scores of Table 2b (0.53 vs 0.79) are more prominent than those in Table 2a (0.60 vs 0.75). This suggests that the exclusion of empty spaces helps to further differentiate models performance.

6 Conclusion

We presented a framework that classifies H&E stained breast cancer WSIs into regions of: normal tissue, benign lesion, in-situ carcinoma and invasive carcinoma. The framework made use of Inception-Resnet-v2 as the underlying patch classifier. The training employed a two-stage approach to utilize both the microscopy images and WSIs. In the first stage, patch classifier was trained using microscopy images. In the second stage, hard examples were extracted from WSIs and the patch classifier was retrained. Prediction results were aggregated from patch-wise predictions back onto image-wise prediction and WSI annotations. Analysis shown that the bias in the evaluation metric deserves much attention. The prediction outcomes are satisfactory but has room for improvement, especially in the detection of in-situ carcinoma and small regions of invasive carcinoma.

This framework won the first place in both Part A and Part B of ICIAR 2018 Grand Challenge on Breast Cancer Histology Images. It achieved a Part A accuracy of 87% which is a 9% improvement over the state-of-the-art. And a Part B score of 0.6929 which outperformed the second place (0.5527) by a large margin.

Acknowledgements. We would like to thank the organizers of ICIAR2018 and BACH2018 who supported and organized this challenge.

References

1. Camelyon16 (2016). https://camelyon16.grand-challenge.org/results/
2. Camelyon17 (2017). https://camelyon17.grand-challenge.org/results/
3. Breast Cancer Facts and Figures 2017–2018 (2018). https://www.cancer.org/research/cancer-facts-statistics/breast-cancer-facts-figures.html
4. ICIAR 2018 Grand Challenge on Breast Cancer Histology Images (2018). https://iciar2018-challenge.grand-challenge.org/
5. Araújo, T., Aresta, G., Castro, E., Rouco, J., Aguiar, P., Eloy, C., Polónia, A., Campilho, A.: Classification of breast cancer histology images using convolutional neural networks. PLOS ONE **12**(6), 1–14 (2017). https://doi.org/10.1371/journal.pone.0177544
6. Elmore, J.G., Longton, G.M., Carney, P.A., Geller, B.M., Onega, T., Tosteson, A.N.A., et al.: Diagnostic concordance among pathologists interpreting breast biopsy specimens. JAMA **313**(11), 1122–1132 (2015). https://doi.org/10.1001/jama.2015.1405
7. Habibzadeh, M.N., Jannesary, M., Aboulkheyr, H., Khosravi, P., Elemento, O., Totonchi, M., Hajirasouliha, I.: Breast cancer histopathological image classification: a deep learning approach. bioRxiv (2018). https://www.biorxiv.org/content/early/2018/01/04/242818
8. Jain, R.K., Mehta, R., Dimitrov, R., Larsson, L.G., Musto, P.M., Hodges, K.B., Ulbright, T.M., Hattab, E.M., Agaram, N., Idrees, M.T., Badve, S.: Atypical ductal hyperplasia: interobserver and intraobserver variability. Mod. Pathol. **24**, 917–923 (2011)
9. Janowczyk, A., Madabhushi, A.: Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases. J. Pathol. Inform. 7(1), 29 (2016). http://www.jpathinformatics.org/article.asp?issn=2153-3539;year=2016;volume=7;issue=1;spage=29;epage=29;aulast=Janowczyk;t=6
10. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. **9**(1), 62–66 (1979)
11. Schnitt, S., Connolly, J., Tavassoli, F.A., Fechner, R., Kempson, R.L., Gelman, R., Page, D.: Interobserver reproducibility in the diagnosis of ductal proliferative breast lesions using standardized criteria. Am. J. Surg. Pathol. **16**(12), 1133–1143 (1992)
12. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. ArXiv e-prints, September 2014
13. Spanhol, F.A., Oliveira, L.S., Petitjean, C., Heutte, L.: A dataset for breast cancer histopathological image classification. IEEE Trans. Biomed. Eng. **63**(7), 1455–1462 (2016)
14. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning, ArXiv e-prints, February 2016
15. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going Deeper with Convolutions. ArXiv e-prints, September 2014
16. Zhong, A., Li, Q.: HMS-MGH-CCDS Camelyon17 presentation (2017). https://camelyon17.grand-challenge.org/serve/public_html/presentations/HMS-MGH-CCDS_Camelyon17_presentation.pptx